Computerized Adaptive Testing and *No Child Left Behind*

G. Gage Kingsbury

Carl Hauser

Northwest Evaluation Association

April 13, 2004

A paper for the session:
Exploration of Pertinent Assessment Issues in Large Scale Testing Environments

Presented at the 2004 Annual Meeting of the American Educational Research Association, San Diego, CA

Computerized Adaptive Testing and *No Child Left Behind*

Among the changes in education called for under the No Child Left Behind act is the need for states to test students in a number of grades and subject areas. Scores from these tests are to be used for a variety of purposes, from identifying whether individual students are proficient, to helping determine whether schools are causing adequate growth for their students. This study investigates several testing approaches, and their potential impact on the use of test scores for these varied purposes.

For the past century, K-12 achievement testing has had a single paradigm. Tests were constructed to assess major content domains according to a collective understanding of how the domain is parceled out into grade-specific units. With rare exception, all students within a grade were administered one test designed for that grade level. The advantages of this **fixed-form** paradigm include its ease of use, and strict control of the content seen by each student.

Over the past two decades, a competing paradigm has emerged in K-12 education. Adaptive testing, which has been used very successfully in the military (Sands, Waters, & McBride, 1997) and in professional certification and licensure (Zara, 1992) has found its way into elementary and secondary education (Kingsbury, 1986). In this paradigm, the test adapts to match the difficulty of the questions administered to the performance of each student as the student takes the test. The advantages of this **adaptive-testing** paradigm include increased testing efficiency, and tests that are challenging but not frustrating for each student (Weiss, 1982).

An example of the adaptive-testing paradigm is seen in the accountability plan that was approved by the federal government for use within the state of Oregon (Tindal and Haladyna, 2002). Oregon has developed tests in a number of distinct levels that are designed with graduated difficulty. Each student takes the test level that is most consistent

with the student's previous classroom performance. This provides a system with accurate measurement across a broad range of student performance. Oregon has recently introduced tests that are to be dynamically administered via computer and have all of the measurement characteristics of an adaptive test.

In an adaptive test, items are selected for administration from a large pool of test questions. The difficulty of test items presented to the student depends on the student's performance on previously presented test items. Higher performance is followed by more difficult questions and lower performance is followed by less difficult questions. The object of the item selection algorithm used to administer test items is to add as much precision as possible to the estimate of the student's achievement. As the test progresses, the estimate becomes increasingly more precise by virtue of continually providing test items that are closer and closer to the student's true achievement level.

One concern involving the use of adaptive testing for NCLB purposes is the requirement that all students to be tested with material that specifically addresses content standards for the grade in which they are enrolled. Modern procedures for selecting items in adaptive tests (Kingsbury & Zara, 1991; Stocking & Swanson, 1993; Van der Linden & Pashley, 2000) have been developed with just this type of application in mind. With these approaches, a variety of constraints can be set on the selection of specific items for administration during the course of an adaptive test. The use of one of these constrained item selection procedures ensures that all items selected in an adaptive test will be appropriate for the content standards required.

Several uses for test scores can be identified with NCLB and its surrounding regulations:

- The most visible use is to identify proficiency categories for students, and to use this information to help schools meet the accountability demands of the legislation.
- The second use of test scores is to identify achievement growth, generally considered as improvement in the percentage of students in a grade level who are considered proficient or above. *Individual* student growth, while not mentioned

explicitly, is perhaps a more interesting if not more important purpose since it pertains to the *C* in NCLB.

- Finally, test results should be expected to inform instruction. In order to enable teachers to move all students forward, information about specific strengths and weaknesses of each student in a class is required

When test scores are to be used for a variety of purposes, the accuracy of the scores for students of different achievement levels becomes a primary concern. Test information functions, which detail the accuracy of scores for a particular test, provide important evidence to help assess how appropriate a test can be for a specific purpose and for a particular set of students. .

Each of the NCLB purposes for test scores has a different implication for the form that the test's information function should take. Along with test content and difficulty, distribution of test information should be a prime consideration in its design. Samejima (1977) identified procedures for connecting the amount of test information in a test with the purposes for which the test was designed. Since that time, the test information function has served as a valuable tool for test design, enabling developers to understand the measurement properties of a test as it was developed.

Figure 1 demonstrates the desired information characteristics of three tests developed to serve different purposes. Panel A depicts an optimal information design for a test designed to make a single-point decision at a score of 200 (e.g., proficient or not). This requires a substantial amount of information focused at the decision point. A test designed to measure student growth is depicted in Panel B. This type of test must provide consistently high levels of information across the entire range of performance. Panel C shows a test designed to identify and support students with special needs (including gifted children and children at risk). This type of test needs high information values at the extremes of the achievement distribution. Perhaps the most important conclusion that can be drawn from Figure 1 is that a test designed optimally for one purpose may not be suitable for all three purposes.

Purpose

The purpose of this demonstration is to use live student data and test information functions to demonstrate the measurement characteristics of two types of tests and identify their utility for NCLB. The two test types are:

- **Fixed form** – This is a single test form, designed to be administered to all students in a particular grade.
- **Adaptive test** – This is a test drawn from an item pool that matches item difficulty to the performance of the student.

This paper demonstrates how test information differs according to test structure and how tests are targeted to students taking them. Along with a basic comparison of test information, the demonstration details the potential impact of using either type of test on our knowledge concerning the students and their school.

Method

Measurement model. All tests and items were calibrated using the one-parameter logistic IRT model (Lord and Novick 1968), also known as the Rasch model (Wright, 1977). This model has the characteristic of being straightforward to implement, while providing the user with the properties of sample independence and scale stability that are required for high-quality measurement.

Tests. Four sets of fixed-form tests, (2 grade levels, 4^{th} and 8^{th} X 2 content areas, Reading and Mathematics) were used as base tests for comparisons. Within each set were two tests of different difficulty, one centered at the 35^{th} percentile and one centered at the 70^{th} percentile. Item difficulties are expressed in RIT values which are simply a linear transformation of the theta metric from the calibration process.

Each fixed-form test had the same general characteristics. For Reading, 40 items were selected to match specific grade level content standards. For Mathematics, 50 items were selected to match specific grade level content standards. Item difficulties were selected to correspond to the classic design of a wide range fixed-form test. This design calls for 36% of the items with difficulties between the mean and 1 sd, 9 percent of the items between 1 sd and 2 sd, and 5 percent of the items between 2 sd and 3 sd. The same percentages would be used for standard deviations below the mean. Because all items had item difficulties derived from applying the Rasch model, a scale score (RIT) and standard error of measurement (SEM) could be determined for each raw score.

Due to their dynamic nature, it is more convenient to examine adaptive tests using empirical test information. To accomplish this 424,328 and 251,399 adaptive Reading test records for grades 4 and 8, respectively were retrieved from the spring 2003 testing season. Adaptive Mathematics test records (428,661 and 368,441 for grades 4 and 8, respectively) were also retrieved from the same testing season. For each RIT score on each test, the mean of the standard error of measurement was calculated.

Information Analysis. The focus of analysis was the level of information yielded by each test across the range of student performance. IRT models have the desirable feature of providing an estimate of the measurement error for each scale score (Baker, 2001). The reciprocal of the squared measurement error for a particular scale score yields the amount of information associated with that score. For both adaptive and fixed-form tests, the calculation of the test information was done in the same manner. However, since the items in an adaptive test differ from student to student, the test information values for the adaptive tests were averaged across all students in the sample, for each final test score.

Impact Analysis. The information functions provide an excellent method to describe the difference between tests, but since they are population independent, they don't fully describe the impact that differences in information might have on a specific group of students. To measure this impact, a simple statistic is developed using the relationship

between the standard deviation of achievement in the population and the standard error of individual test scores. The development of the statistic goes as follows:

- Before we give a test to a student, the best estimate of the student's achievement is the mean achievement level in the population. The standard error of this estimate is equal to the standard deviation of achievement in the population.
- When a test has been administered to a student, we have gained information to improve our estimate of the student's achievement level. The standard error of the score is reduced in a fixed relationship to the amount of information we have gained.
- The ratio of the standard error to the standard deviation (the **standard ratio**) indicates the degree to which we have reduced our uncertainty about the student's achievement level. The ratio shrinks from 1 (when the standard error equals the standard deviation, prior to testing) as we add information about the student.
- If we can shrink the ratio to .30 or less, our test provides us substantial information about the student, and this information can be used to make meaningful instructional decisions concerning the student. If we can not achieve this ratio or better, the test scores will be substantially less useful for making fine distinctions among students and their respective needs.

For this study we calculate the percentage of students for which each test is not able to meet the ratio of .30 or better. This serves as a direct indication of the percentage of students for whom instructionally effective information may not be available.

It should be noted that other studies looking at differences among tests have commonly used the relative efficiency measure (Lord, 1980). While that statistic has its uses, it isn't very good for identifying the impact of differences in precision on student scores. As an example, a test that provides extremely little information across a particular achievement range would have a high relative efficiency of 2.0 in that range if the comparison test provided information that was half of "extremely little". Even with the substantial difference indicated by the relative efficiency, both tests would still be poor measures for the students in that particular achievement range .

Demonstration

Information. Figure 2 presents the information functions for all the tests examined. The figure is divided into four panels, A-D. Each panel is made up of two charts. The top chart presents the test information functions for the particular grade level and content area, and lower chart presents a distribution of student achievement. The achievement distributions are presented to provide a context for the test information functions. Within each chart of a set, a few details are important to understand.

The top chart in each set provides three test information functions, one for each test type. The lower two information functions, appearing as somewhat normally distributed, are for the two fix-form tests. Of these two, the function depicted in a solid line is for the test centered at the 35th percentile; the one depicted as a dashed line is for the test centered at the 70th percentile. The information function running mostly across the top of the chart is for the adaptive test. In each of the information function charts , the lowest and highest RIT scores for the subject are set at approximately percentiles 1 and 99. The vertical line in the center of the charts represents the cut score points for proficient performance.

The results in Figure 2 show that in each comparison the adaptive test provides the most information at all achievement levels. This finding is in keeping with virtually all earlier research comparing adaptive testing to fixed-form testing. What is not so expected is the magnitude of the difference in information for students in the extremes of the achievement distribution. For students in the lower end of the achievement range (below 170 on the measurement scale), the adaptive test provides more than three times the information provided by the fixed form test. For these students, the use of a fixed-form test will deprive teachers of needed information.

Impact. The results of the standard ratio analysis are presented in Table 1. This table shows the percentage of students whose scores would be expected to exceed a ratio of .3. These results are based on the use of population standard deviations from a norming study

with over 70,000 students per grade level. With both of the conventional fixed-form tests in each subject and grade level combination, portions of the expected student scores that exceed the criterion (hence yielding minimal information) are large enough to warrant questioning the use of these tests for these populations.

It should be noted that the selection of a criterion is a matter of informed professional judgment. In this example, if the criterion level for the standard ratio had been changed from .3 to .25, the results for the adaptive tests would remain virtually unchanged while close to 100% of the expected student scores on the fixed-form tests would fail to reach the desired impact ratio.

Discussion

The varied purposes for assessment information mandated by NCLB require test designs with a substantial amount of psychometric sophistication. A fixed-form test that is most appropriate for making decisions concerning the placement of students into categories will not be appropriate for making instructional decisions, or for measuring student growth. In order to fulfill all of the purposes of NCLB, an assessment must be accurate for all students. While a very long, fixed-form test with a wide range of item difficulties might meet this need from a psychometric standpoint, it would have substantial drawbacks, including the following:

- It would require a vast expenditure of student time
- It would be a psychologically unappealing test, since students would be bored by the easy content, and frustrated by the difficult content
- It would waste much of the time spent, since test questions that aren't at the point of challenge for a student provide little information about the student's capabilities

From the demonstration above, it can be seen that an adaptive testing solution provides an information function that more closely approximates the information needed to meet all requirements of NCLB. In the analysis of information functions, the adaptive test provided more information at every level of achievement than either fixed-form test. While some drop-off in information was noted for the highest and lowest performing students for the adaptive test, the test provided consistently higher information than the fixed-form tests.

The results from the impact analysis identify the percentage of students who aren't being measured with the precision required for instructional decision making. In no condition did the adaptive test have *more than* 1 percent of the students imprecisely measured. In no condition did the either fixed-form test have *less than* 6 percent of the students imprecisely measured. This difference in the impact on students' scores may become even more important as the AYP targets become more rigorous for schools.

In addition to the psychometric issues, however, the use of adaptive testing would also have the following advantages for use with NCLB:

- Students would be challenged, but not frustrated by each question on the assessment.
- Students' scores would be as accurate as possible, for any given test length.
- Students, teachers, and other stakeholders could receive immediate feedback concerning student, class, and school performance.

As NCLB moves forward in implementation, it is crucial that educational agencies have precise information about the achievement of their students. As this study has shown, a transition from fixed-form testing to adaptive testing may help agencies with their information needs, without adding unduly to the amount of testing being done in classrooms. The use of adaptive tests may also provide teachers with information that is timely and useful for instruction. If we expect to move all children ahead, we need to know where they are today. Adaptive testing is one approach to providing that information.

References

Baker, F. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Kingsbury, G. G. (1986). Computerized adaptive testing: A pilot project. In W. C. Ryan (Ed.), Proceedings: NECC '86, National Educational Computing Conference. Eugene, OR: University of Oregon, International Council on Computers in Education.

Kingsbury, G. & Zara, A. (1991). A Comparison of Procedures for Content-Sensitive Item Selection in Computerized Adaptive Tests. *Applied Measurement in Education*, 4 (3) 241-261.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233-247.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing : From inquiry to operation*. Washington, DC: APA Books.

Stocking, M.L. & Swanson, L. (1993). A Method for Severely Constrained Item Selection in Adaptive Testing. *Applied Psychological Measurement*, 17(3), 277-292.

Tindal, G. & Haladyna, T. M. (Eds.)(2002). *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. Hillsdale, NJ: Erlbaum.

Van der Linden, W. J. & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (eds.), *Computerized adaptive testing: Theory and practice* (pp.1-25). Boston: Kluwer.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Zara, A. R. (April, 1992). *A comparison of computerized adaptive and paper-and-pencil versions of the national registered nurse licensure examination*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Figure 1. Test information functions.

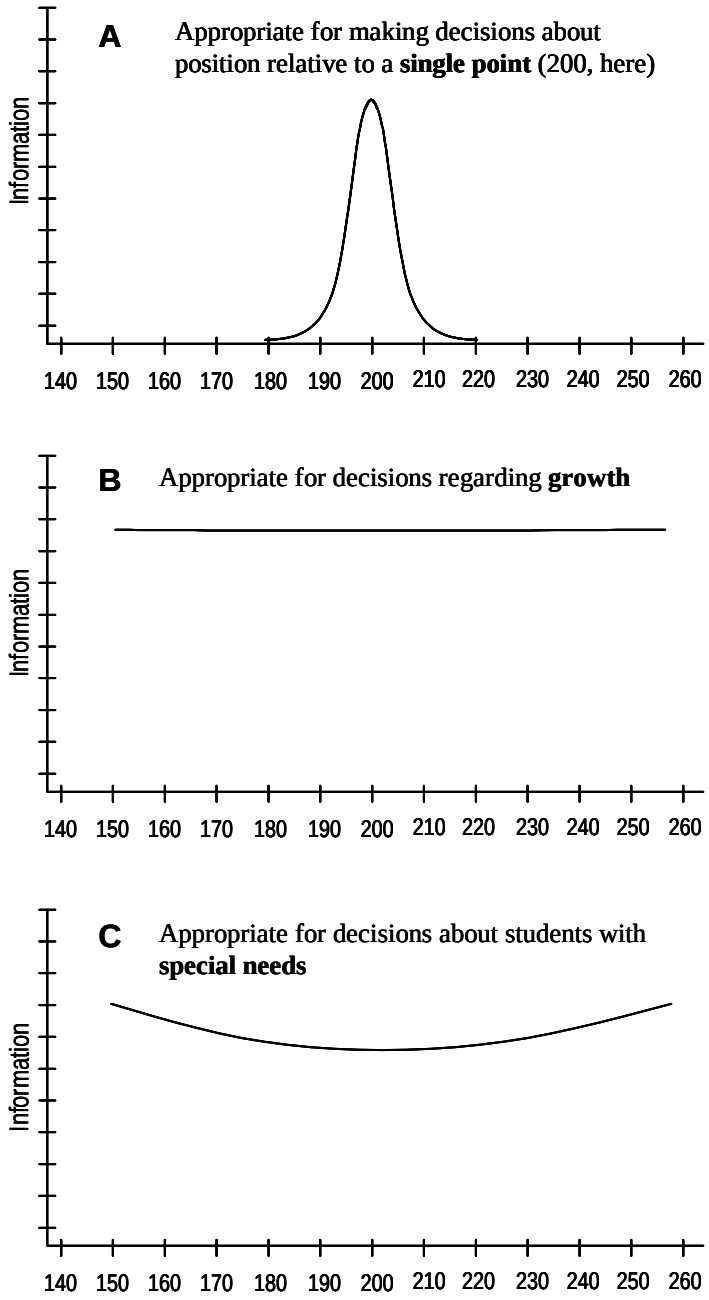

Figure 2. Test information functions for adaptive and fixed-form tests in Reading and Mathematics in grades 4 and 8.

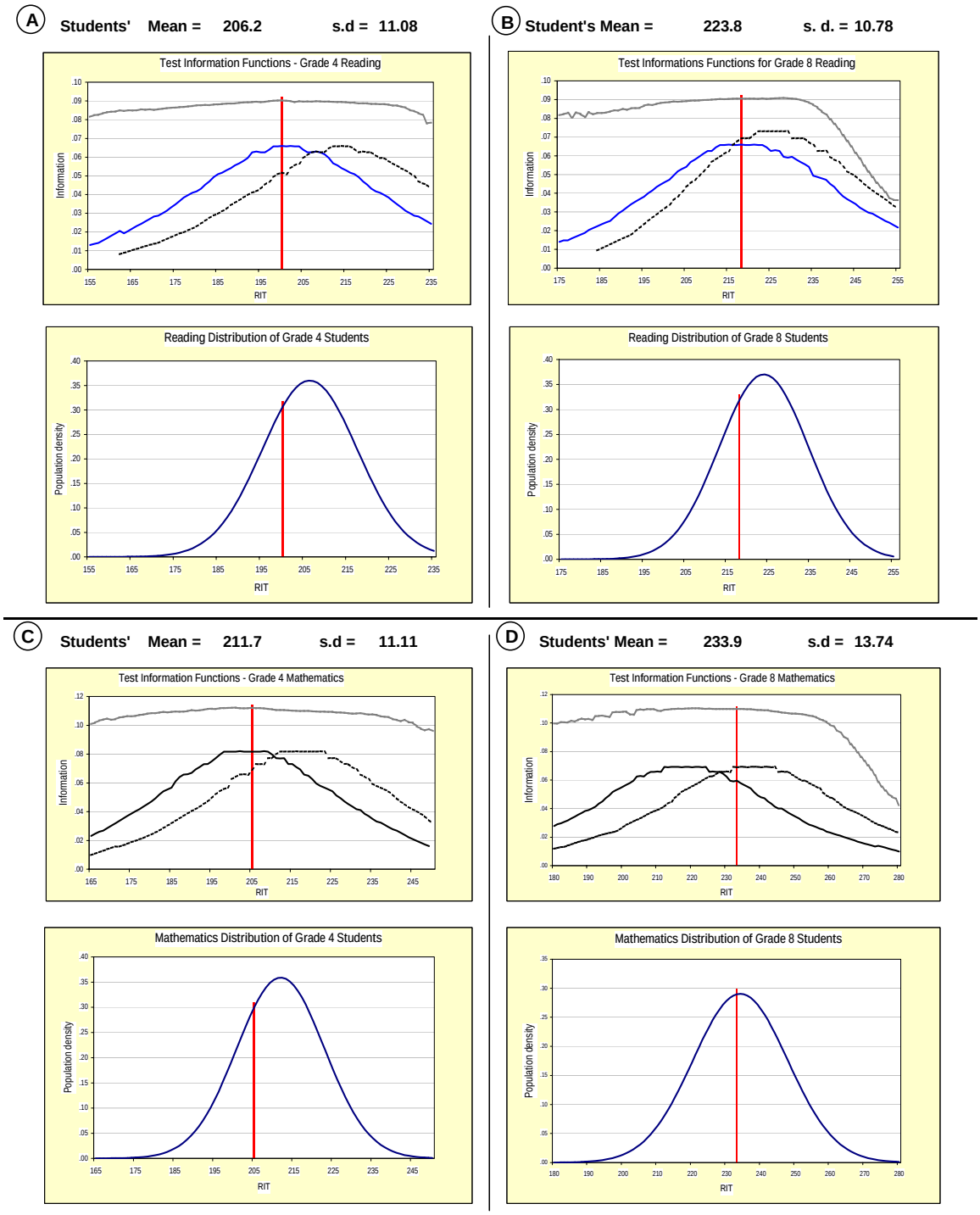

Table 1. Approximate percentages of students whose test scores would be associated with minimal information.*

	READING			MATHEMATICS		
		Fixed Form			Fixed Form	
Grade	Adaptive	Moderate	Hard	Adaptive	Moderate	Hard
4	0.0	25.7	31.9	0.0	20.4	27.3
8	0.0	29.3	26.0	0.8	15.7	6.6

* Minimal information was defined as the standard error of measurement >= .3 standard deviations from the NWEA 2002 grade level norms.